DERIVATIVES
China | 衍盛中国

Machine Learning Methods For Stock Selection

Yang Wang
wangyang@cslt.riit.tsinghua.edu.cn

Introduction

- There are are some systemic risk in the market that are difficult to predict.
- To avoid these risk, we can find some excess returns (alpha returns) and hedge.
  - Long a basket and short a basket.
  - Use financial derivatives
- Some factors contains the information of alpha returns

Multifactor Strategy

- Task : $\{f_1,\ldots f_n\}$ -> $\{s_1,\ldots,s_m\}$, where f_i is the ith factor, s_j is the jth stock.
- Artificial method: Scoring the stocks
- ML method:
  - Regression
  - Classification

Multifactor with ML

- Regression: select the top-k according to the predicted returns
  - Pros:
    - Can describe the returns
  - Cons:
    - Vulnerable to noise
    - Can not describe the confidence

Multifactor with ML

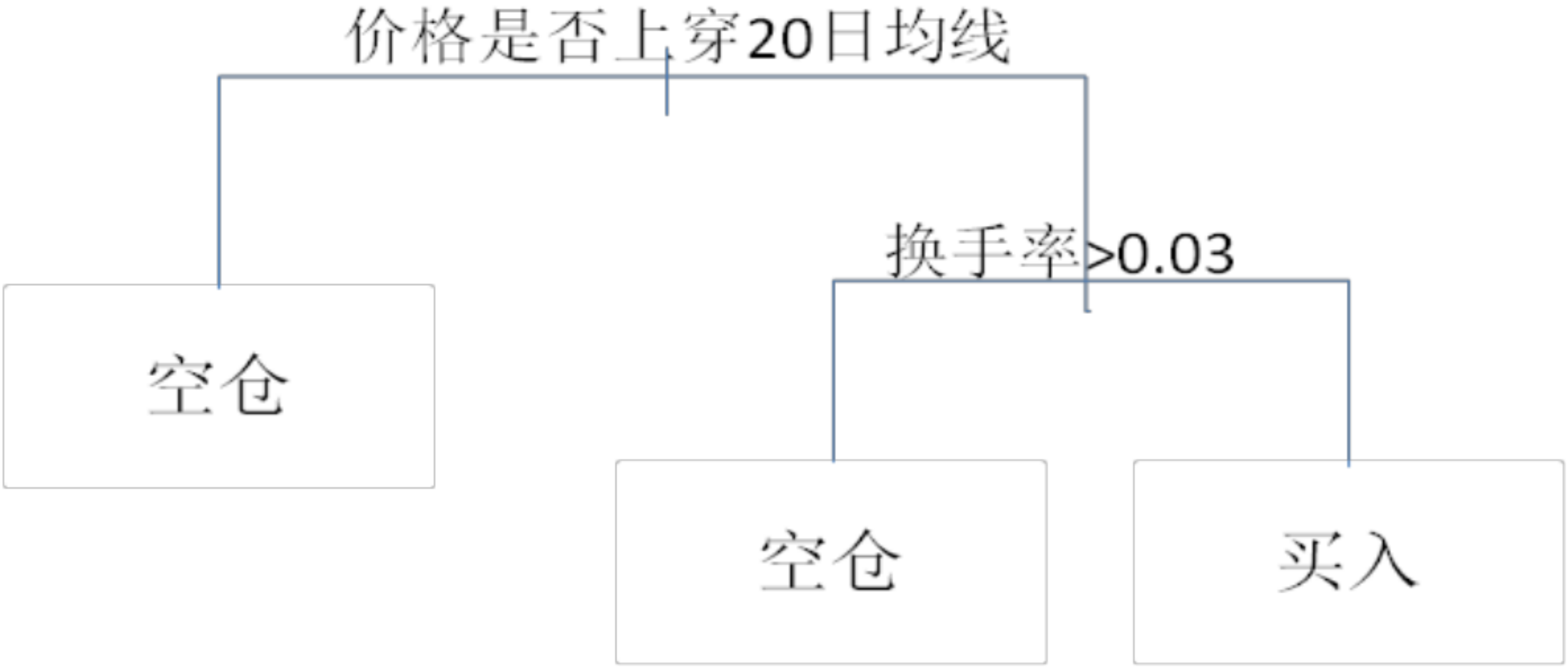
- Classification: select the top-k according to the confidence
  - Pros:
    - Can describe the confidence
    - Robust to noise
  - Cons:
    - Can not describe the returns

Solution: discretize the returns and use multi-classification

Decision Tree

- Motivation :
  - Many investors have no support from profession teams so they are used to trade according to indicator.
  - The process can be described by a tree

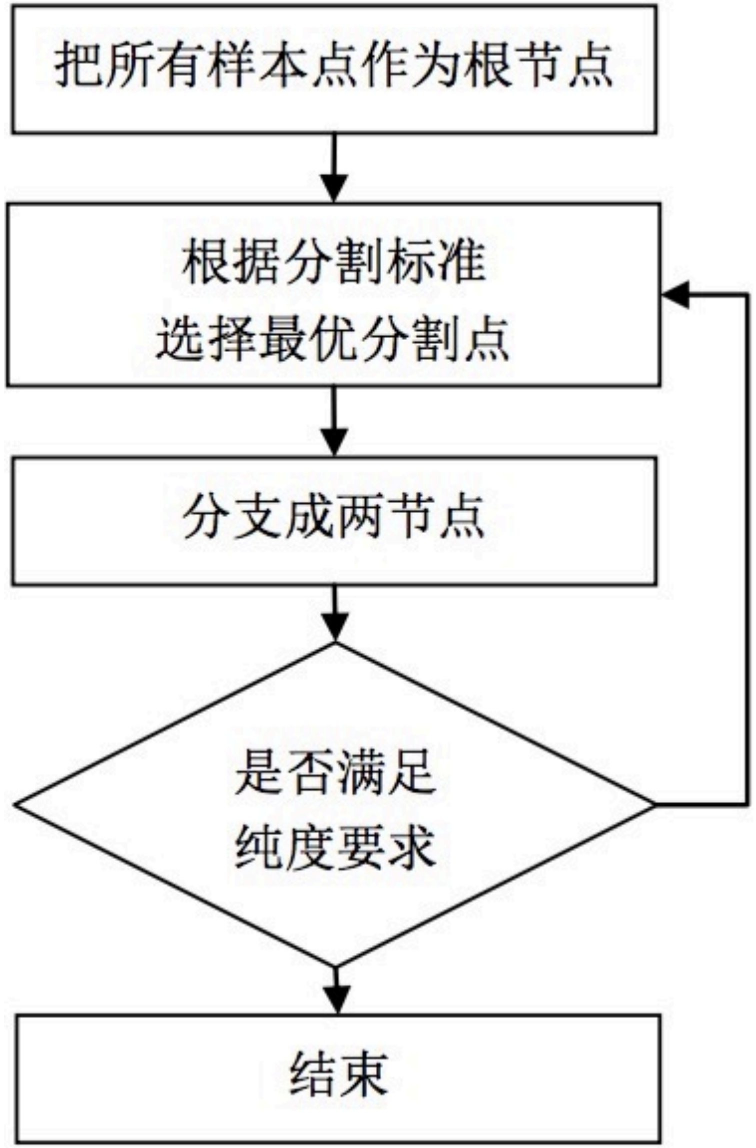
Decision Tree



Decision Tree

- Algorithm
- Gini不纯度

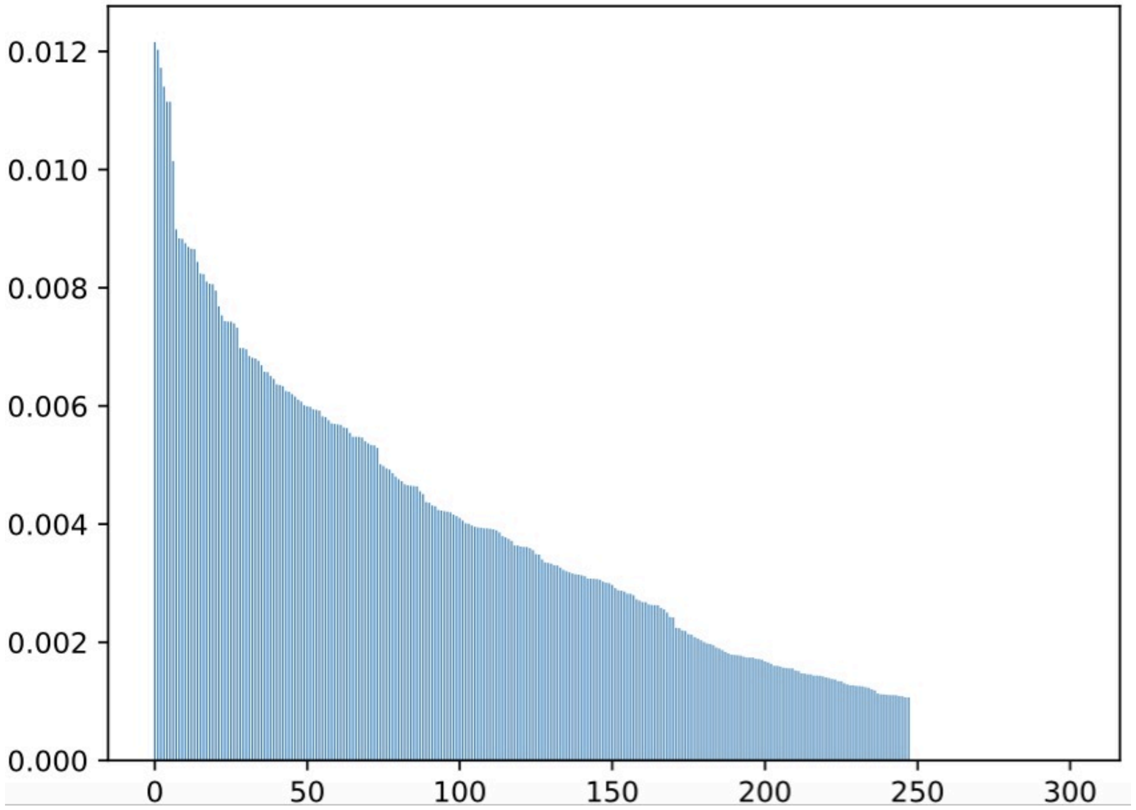
Gini = $\sum_{k=1}^{K} P(m,k)(1-P(m,k))$



Decision Tree

- Feature importance analysis:

$$importance(f_i) = \sum_{\text{node}} gini_{\downarrow}$$

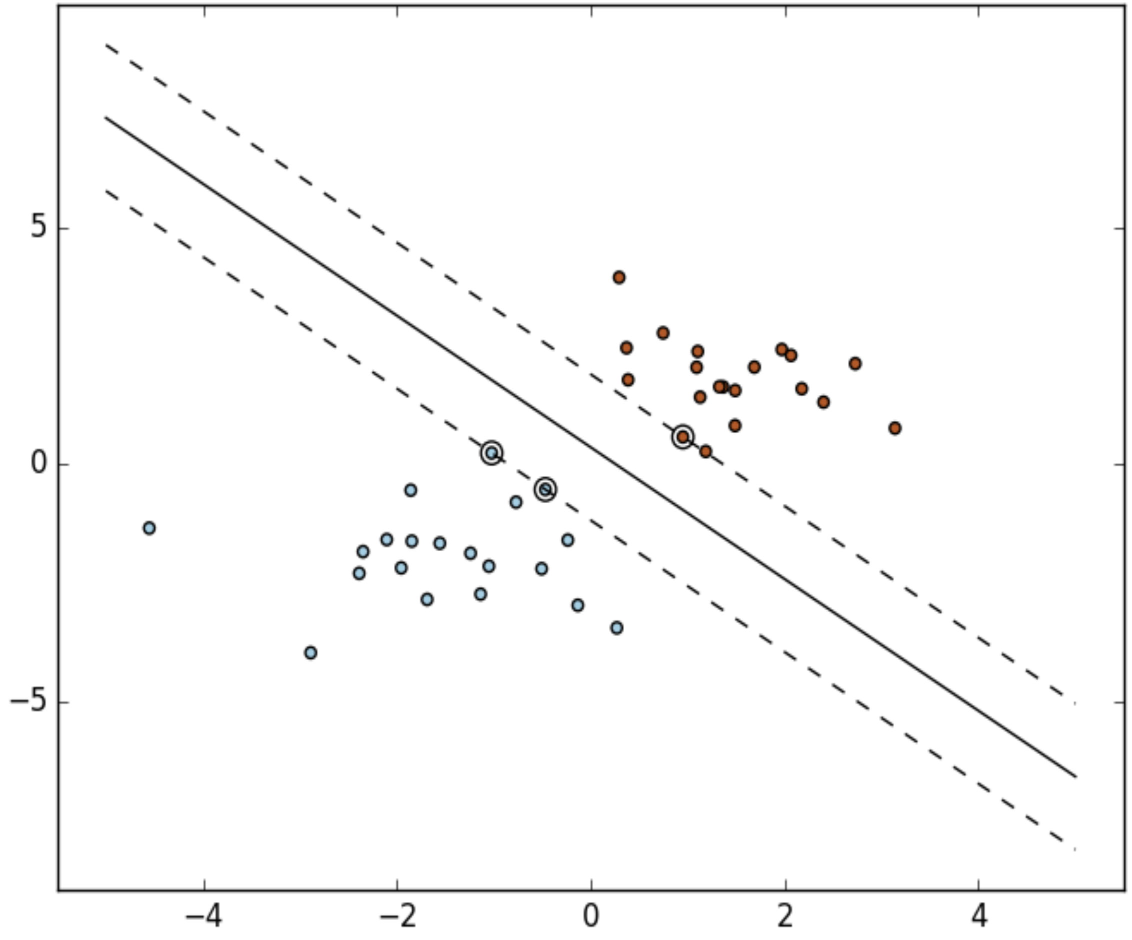


Decision Tree

- Overfitting control:
  - Limit the depth
  - Limit the number of leaf nodes
  - Limit the minimum of examples for splitting
  - Limit the minimum decrease of Gini

SVM

- Goal :
  - Separate the different class points as wide as possible



SVM

- Objective function:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0, i = 1, ..., n$$

Experiments

- Setting:
  - Change position every month
  - Window size for training
    - Decision tree: i-24~i-1 month
    - SVM: i-30~i-1 month
  - Portfolio size (uniformly)
    - CSI300: 20
    - ZZ500: 30
  - Bid price: vwap

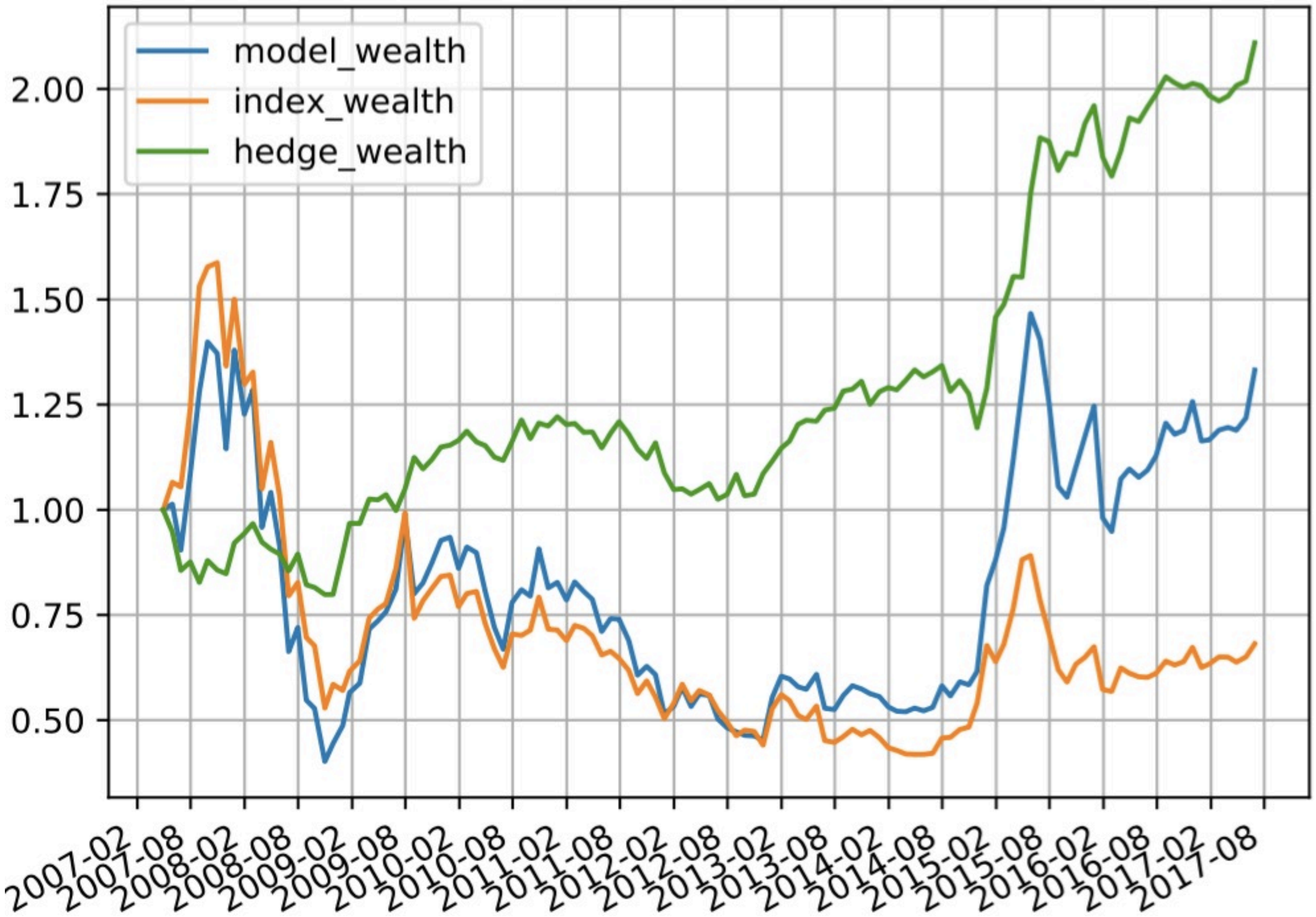
Performance

Profits: 110.98%

Sharpe ratio: 0.62

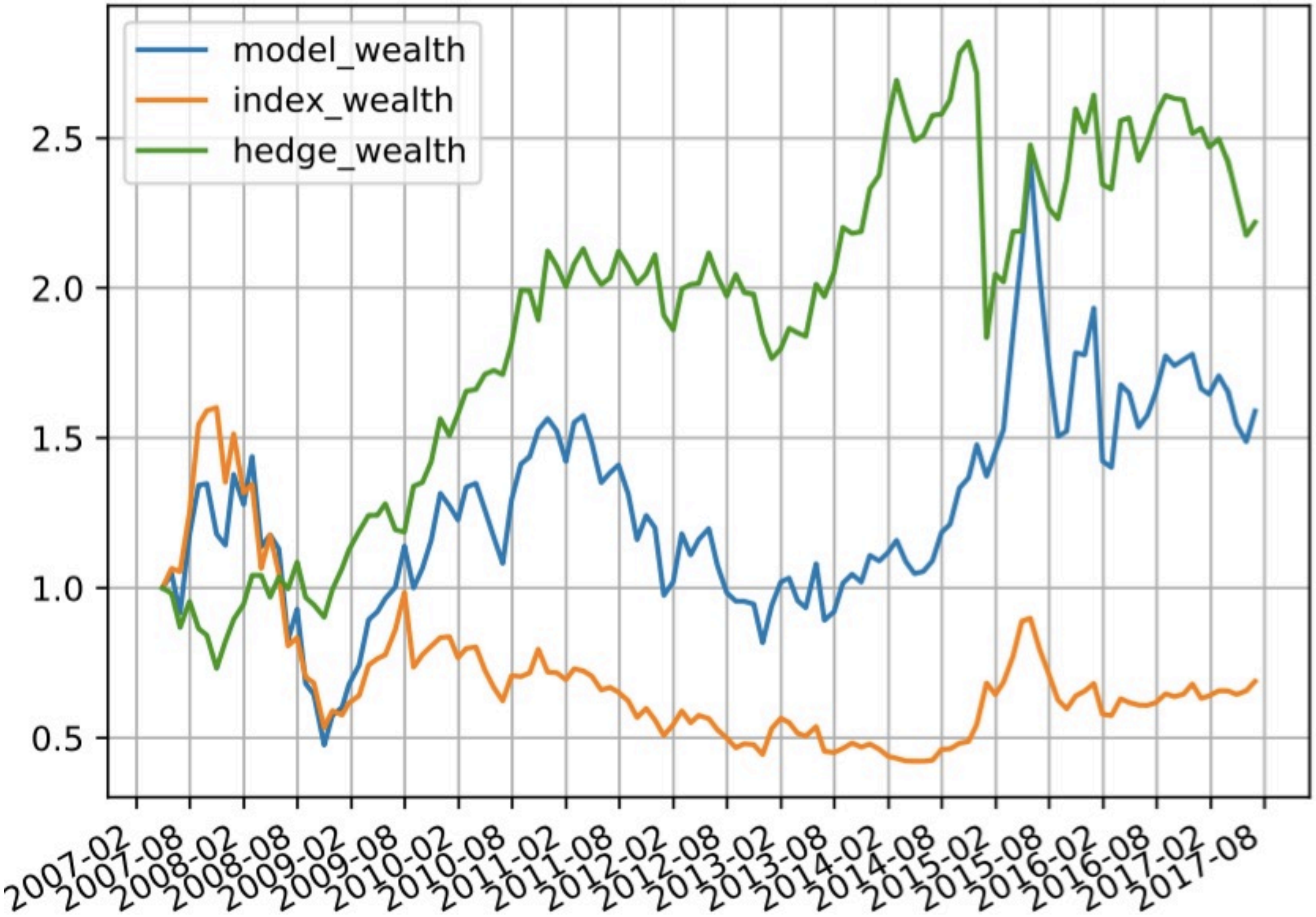
Max drawdown : 20%

Decision Tree (CSI300):



Performance

Decision Tree (ZZ500):



Profits: 121.96%

Sharpe ratio: 0.47

Max drawdown : 35%

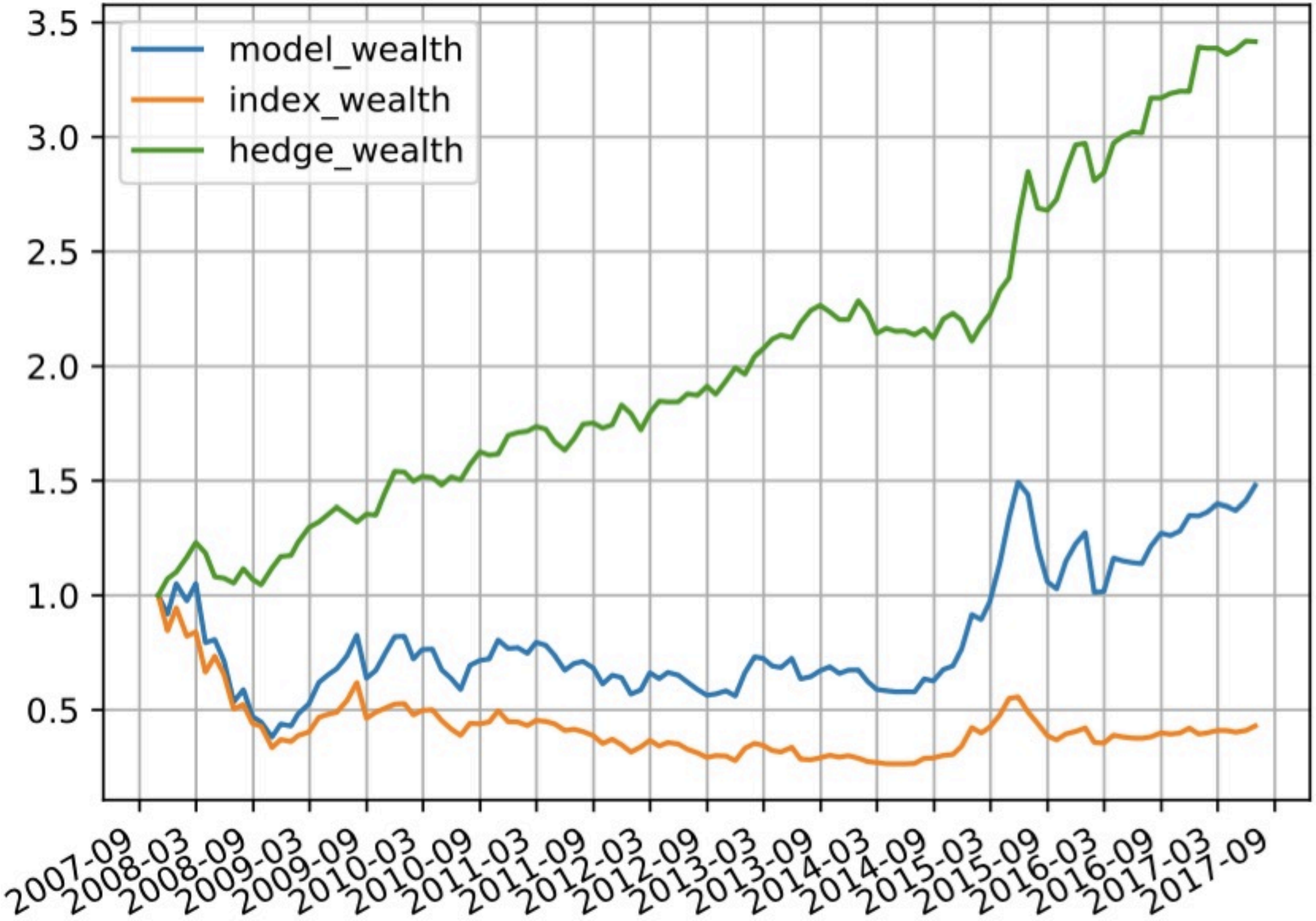
Performance

Profits: 241.65%

Sharpe ratio: 1.22

Max drawdown : 15%

SVM (CSI300):



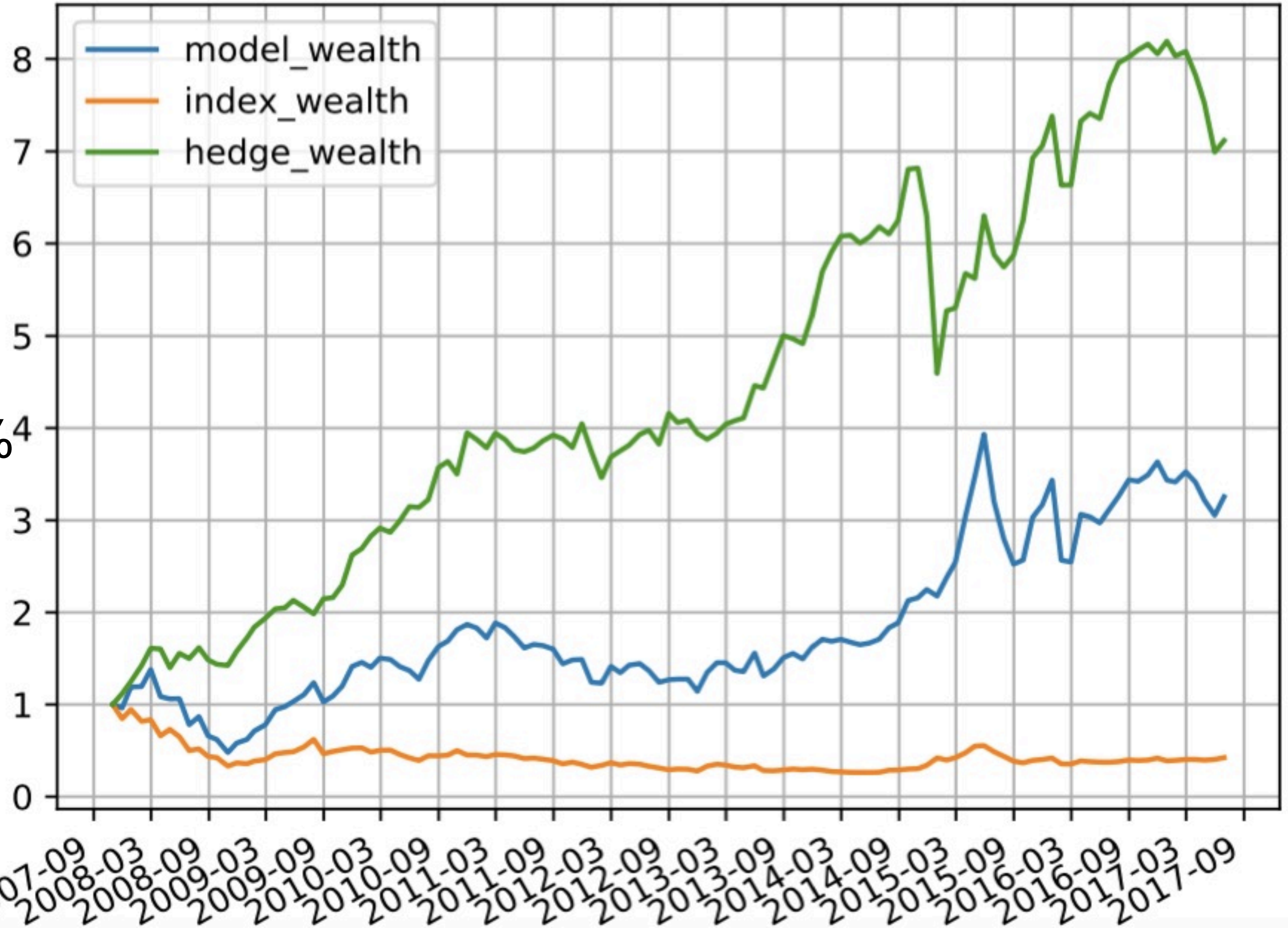
Performance

Profits: 611.82%

Sharpe ratio: 1.08

Max drawdown : 32.64%

SVM (ZZ500):



Conclusion & Future Work

- ML methods can achieve a not bad results.
- SVM is more robust than decision tree for multifactor-based strategy
- CSI300 is more stable and ZZ500 is more profitable
- A more detailed and realistic backtesting need to be done
- Good combination of CSI300 and ZZ500 will be valuable

Thank you for your attention！

DERIVATIVES
China | 衍盛中国

wangyang@cslt.riit.tsinghua.edu.cn